Chapter 9

Generating Training Sets for the Automatic Recognition of Handwritten Documents

Gabriel Pereira e Silva and Rafael Dueire Lins

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772//52074

1. Introduction

Handwritten character recognition is a task of high complexity even for humans, sometimes. People have different writing "style", which may vary according to psychological state, the kind of document written, and even physical elements such as the texture of the paper and kind of pencil or pen used. Despite such wide range of variation possibilities, some elements tend to remain unchanged in a way that other people, in general, can recognize one´s writing and even identify the authorship of a document. Very seldom one is unable to identify one's own writing. Very seldom someone is unable to identify his own writing.

The basis for pattern recognition rests on two corner stones. The first one is to find the minimal set of features that presents all maximum diversity within the universe of study. The second one is to find a suitable training set that also covers all possible data to be classified. Due to the variation of writing styles between people, one should not expect that a general classifier yields good recognition performance in a general context. Thus, one tends to either have general classifiers for very specific restricted vocabularies (such as digits), or to have personalized recognizers for general contexts. The scope of the present work is the latter. In such context it is a burden and very difficult to generate a good training set to allow the classifier to reach a reasonable recognition rate.

This paper proposes a new approach for the automatic generation of the training set for the handwritten recognizer of a given person. The first step for that is to select a set of documents representative of the author's style. In the Internet one may find several public domain sites with font sets. In particular the site Fontspace [21] offers 282 different cursive font sets for download (e.g Brannboll Small, Jenna Sue, Signerica Fat, The Only Exception, Homemade Apple, Santos Dumont, etc). Figure 1 presents an example of some of them. The key idea presented here is "approximating" the author writing by a cursive typographical font, which is skeletonized and a "standard" training set is generated. Such strategy,

detailed as follows, was adopted with success with documents of the Nabuco bequest [12] and of the Thanatos Project [1].

Figure 1. Examples of cursive font sets extracted from Fontspace [1]

2. The proposed method

The choice of a representative training set together with a good set of features is fundamental for the success of automatic pattern recognition. These two factors are tightly linked to each other and in such a way as to grant good recognition results. To obtain a good training set for handwriting recognition of a given author during a period of time in which the writing features are stable (they changes with age, psychological and health factors, etc.) one has to group together the documents that have similar properties. A subset of them that is representative of the set of documents (in general the size of the training set is about 10% of the size of the whole data "universe") is chosen in such a way as to cover the whole diversity of the documents to be transcribed.

The development of the proposed method starts by using a set of cursive fonts. In the Internet one may find several public domain sites with font sets. In particular the site Fontspace [21] offers 282 different cursive font sets for download.

The central difficulty in generating the training set for handwritten documents is to have a "font set" of a specific author to extract the convenient features for patter matching.

The strategy adopted here is:

1. Select a number of documents that is representative of the author.
2. Process the documents to:
   a. remove digitalization borders that may frame the image,
   b. correct skew,
   c. filter-out back-to-front interference (bleeding),
   d. binarize the document.
3. Transcribe the document into text by a human reader.
4. The user should select a number of a cursive font set that bears "some resemblance" to the original author handwriting.
5. The text version of the document is typeset in each of the cursive font sets chosen in step 4 (above).
6. All the typeset versions of the document are converted into image.
7. The image of the original document is skeletonised and then dilated.
8. Segment the image in boxes around each letter (font cases) of the skeletonised and dilated version of the original image and the synthetically generated images.
9. Apply a deformation transform to make each font case in the synthetic images coincide with the font case of the skeletonised-dilated version of the original document.
10. Extract the features from each document and place in a vector.
11. Take the Hamming distance between the feature vectors of the synthetic and original images.
12. The font set used to generate the synthetic document which provides the smallest Hamming distance is the one to be used as the training set.

The structural features used for pattern recognition, mentioned in step 10 above, are:

- Geometric Moments [15] [9];

- Concavity Measurements [16];
- Shape Representation of Profile [14];
- Distance between barycentre points between two consecutive characters;
- Maximum and minimum heights of two consecutive characters.
- Maximum and minimum distance between concavities of two consecutive characters.

The image filtering operations listed in step 2 were performed using HistDoc v.2.0 environment [13] which offers a wide number of tools for historic document image processing including the several algorithms for the removal of back-to-front interference and binarizarion. The skeletonization and dilation processes in this work were performed using the filters available in ImageJ [20].

Step 9, performing image vectorization, is important to increase the likelihood between the synthetic and the original documents. Such operation is applied to each character in each synthetically generated image by deforming the bounding-box and the strokes until there is a perfect match between the synthetic and the original one. In this "deformation" process some statistical analysis is performed to infer data about inter character and inter word spacing, line and character skew, inter line separation, etc.

The feature vector of a document brings an account of the basic features of the author calligraphy. The Hamming distance between the feature vectors of the synthetic and original images, which is part of step 11, brings an account of their similarity, and is calculated using the formula:

$$H_w = \sum_{n=1}^{N\ features} \left| f_{on} - f_{sn} \right|$$

where f_{on} and f_{sn} are the components of the feature vectors of the original and synthetic images, respectively. The choice of a vector of features such that one could extract "information" about the calligraphic pattern of the author shares some ideas with the work in reference [5]. The font set that provides the smallest Hamming distance to the original set is chosen to synthetically generate the whole training dictionary to the classifier.

In what follows the steps described above are detailed in two files of historical documents: the handwritten letters of Joaquim Nabuco that are about one century old and the hand filled information on the books of pre-printed forms of civil certificates from the state of Pernambuco-Brazil, from mid 20th century.

3. Results

The strategy presented above for developing the training set was tested in two sets of documents: letters from Nabuco bequest and death certificates from the Thanatos project [1].

3.1. Transcribing Nabuco´s letters

The Nabuco Project [12] was an initiative of the second author of this paper. It started in 1991 with the aim to preserve the file of letters of Joaquim Nabuco, a Brazilian statesman,

writer, and diplomat, one of the key figures in the campaign for freeing black slaves in Brazil (b.1861-d.1910). The Nabuco file encompasses over 6,500 documents and about 30,000 pages of active and passive correspondence (including postcards, typed and handwritten letters), a bequest of historical documents of paramount importance to understand the formation of the political and social structure of the countries in the Americas and their relationship with other countries. The letters of Nabuco were catalogued and some of them summarized [2] [4], but the bequest was never fully transcribed. The Nabuco Project is acknowledged as being the pioneering initiative in Latin America to attempt to generate a digital library of historic documents. Figure 2 presents an example of letter in Nabuco bequest, written in a blank sheet of paper without lines, which presents a textured background due to paper aging, a horizontal folding mark in its central part, a light back-to-front interference (bleeding) as the letter was written on both sides of the sheet of paper. The image was acquired with an operator driven flatbed scanner, using 200 dpi resolution, in true color. There is no marginal noise (borders) framing the image and its skew is negligible.

To automatically generate the training set for recognizing the handwritten letters from Nabuco file a visual inspection was made to find letters that could represent the whole universe of letters. From the Nabuco file 50 letters were chosen and transcribed by historians, yielding 50 text files, totaling 3,584 words. Twenty-five letters (1,469 words) were used to develop the feature set used for training the classifier, and the remaining ones for ground-truth testing. All the selected documents were processed performing the steps listed in step 2 above, which encompasses marginal border removal, image de-skew, removal of back-to-front interference and binarization. An example of resulting document after filtering and binarization using the HistDoc v.2.0 environment [13] may be found in Figure 3. The image in Figure 2 is skeletonized and then dilated using the filters in ImageJ [20]; the resulting image is presented in Figure 4.

The synthetic image generation is performed by choosing a subset of the cursive fontsets available that resemble the writing of the original document. In the case of Nabuco, the subset selected encompassed 15 of the 282 cursive font sets available in Fontspace [21] (e.g Brannboll Small, Jenna Sue, Signerica Fat, The Only Exception, Homemade Apple, Santos Dumont, etc) that were closer to the author's writing style during the period of interest. The text of the original document was (human) transcribed into a text file, which was typeset using the choosen cursive fontsets. The text of document in Figure 2 typeset using the fonts in "Signerica Fat" type font is shown in Figure 5. The image shown in Figure 5 is now vectorized and "approximated" to the original skeletonized and dilated image by "deforming" each "letter case" and strokes until matching, as much as possible. The resulting image is presented in Figure 6.

The feature vector of each of the synthetic images "deformed" in such a way to the character case to match the original font case was extracted and the Hamming distance of each of them to the skelotonized-dilated original image was calculated. The image that exhibited the minimum distance was the "Signerica Fat" font set presented in Figure 6.

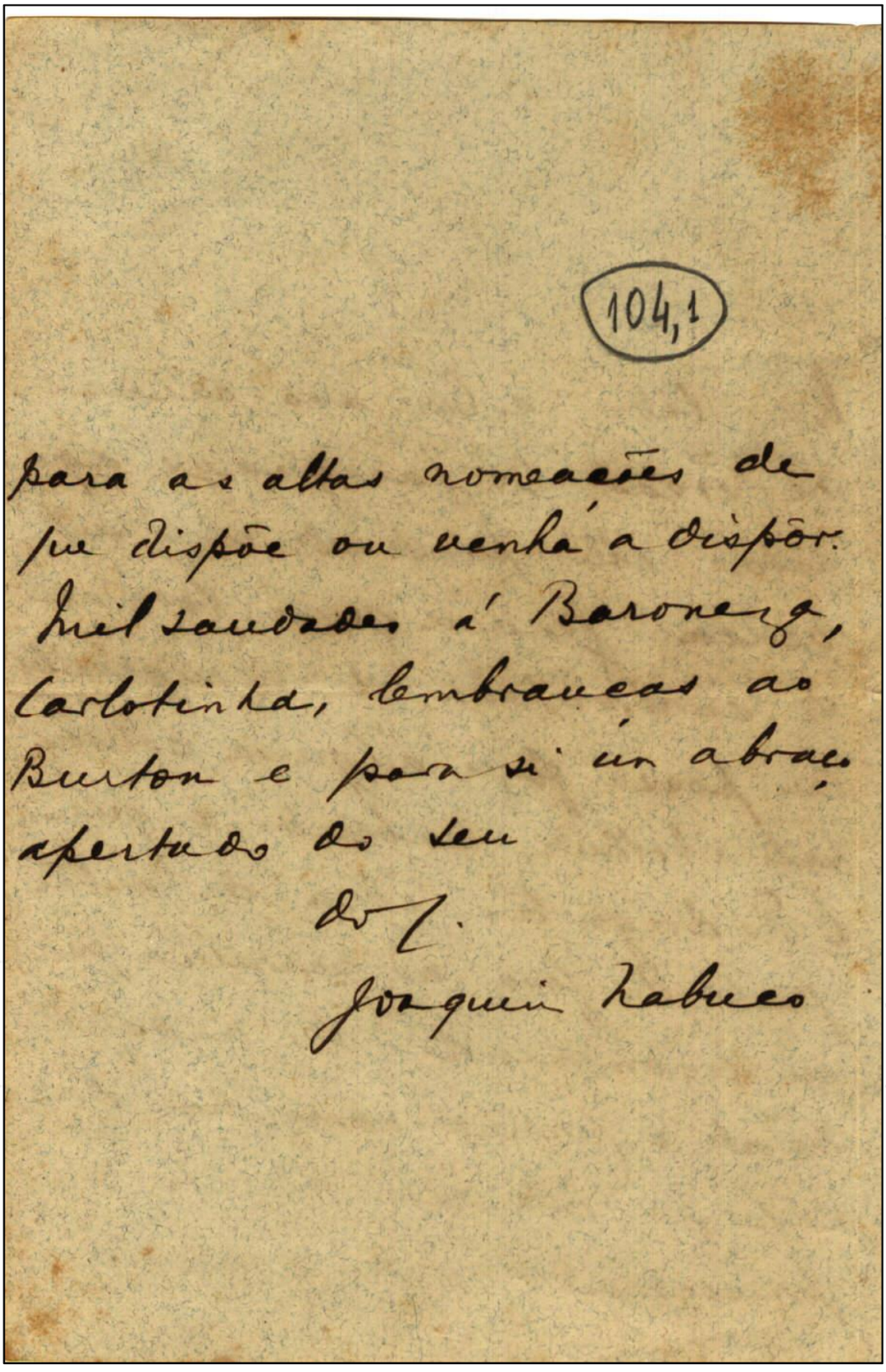


Figure 2. Letter from Nabuco bequest.

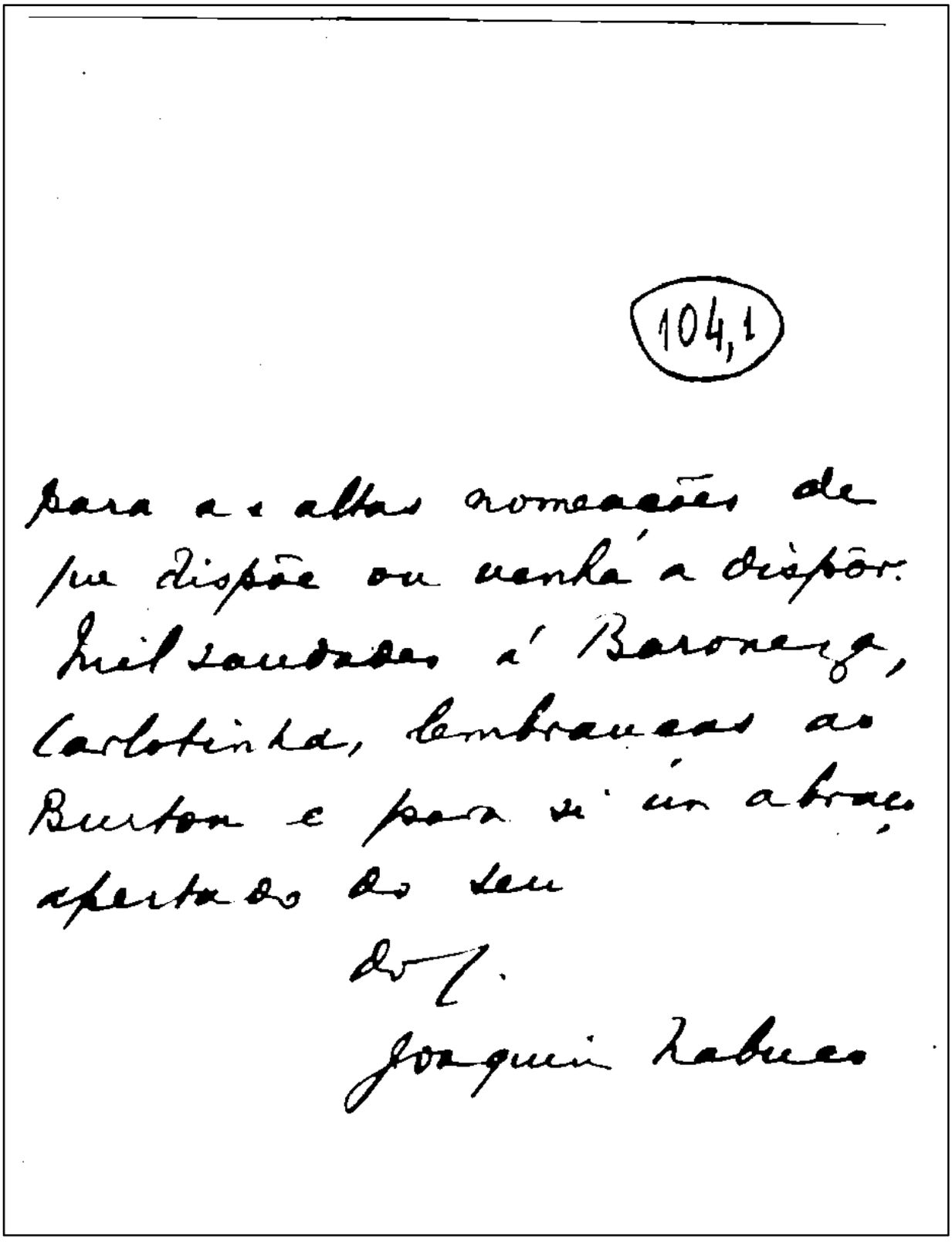


Figure 3. Document shown if Figure 2 after filtering and binarization.

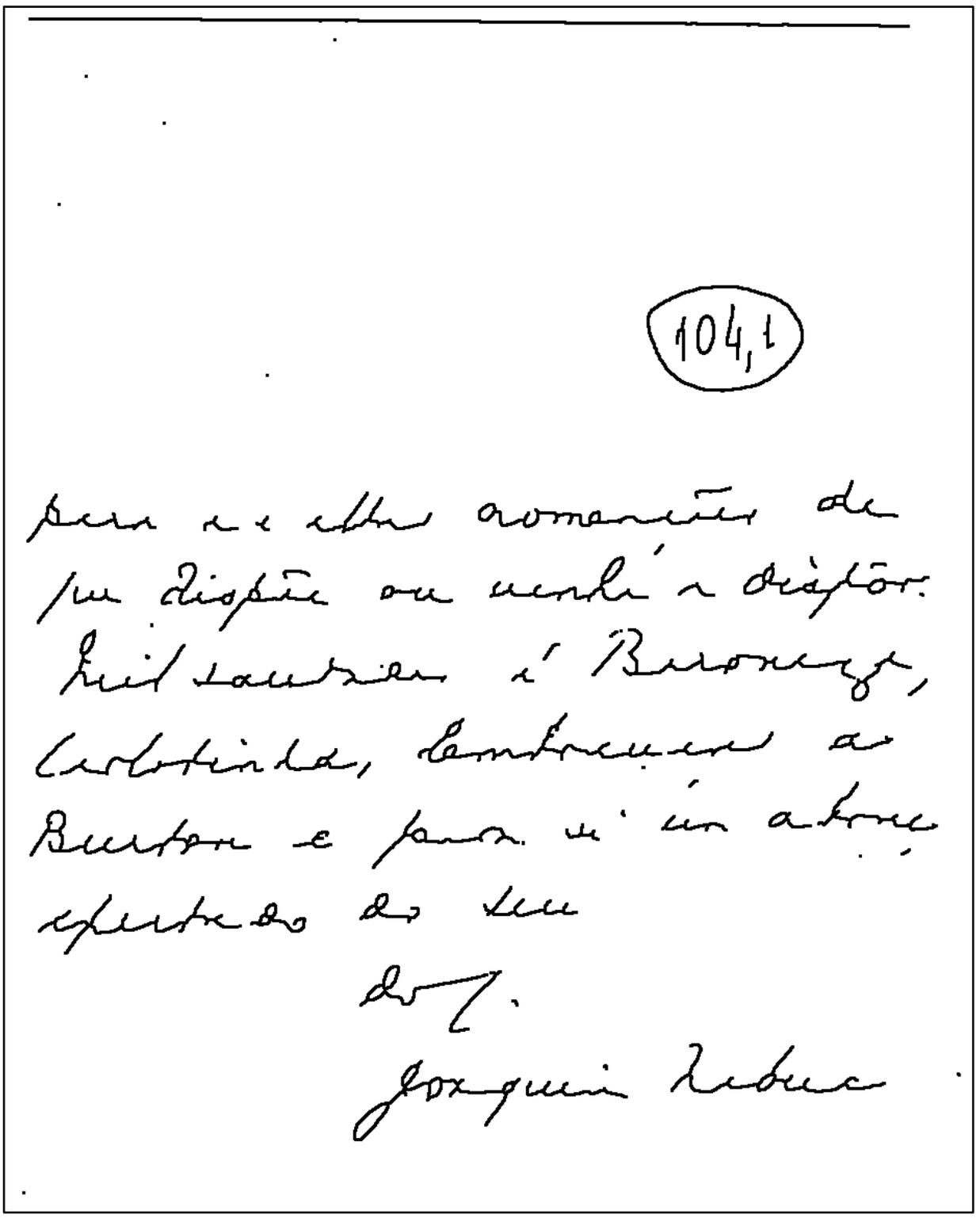


Figure 4. Skeletonized and dilated version of letter shown if Figure 3

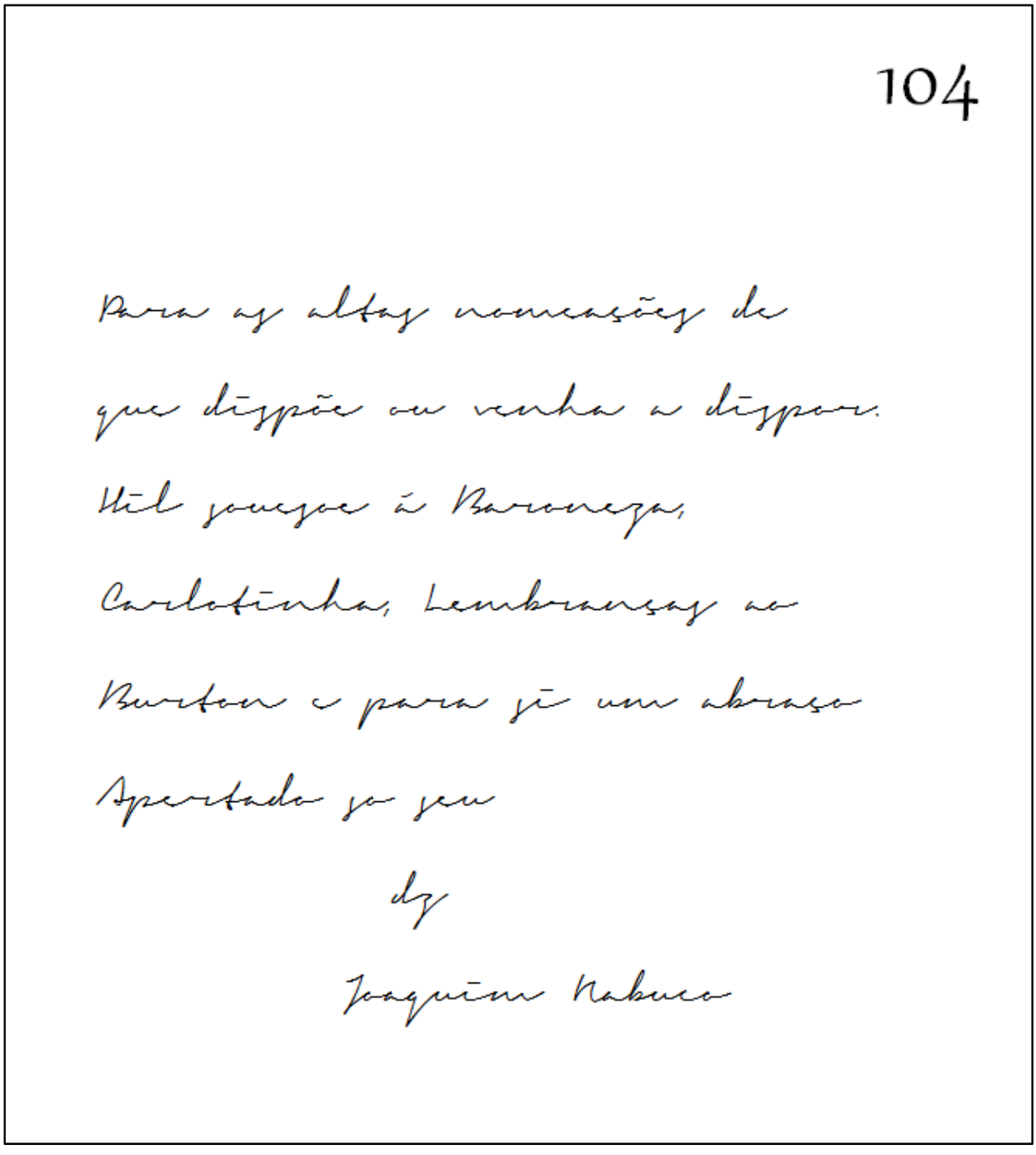


Figure 5. Synthetic skeletonized image generated from typesetting the text of the original document with the "Signerica Fat" font set.

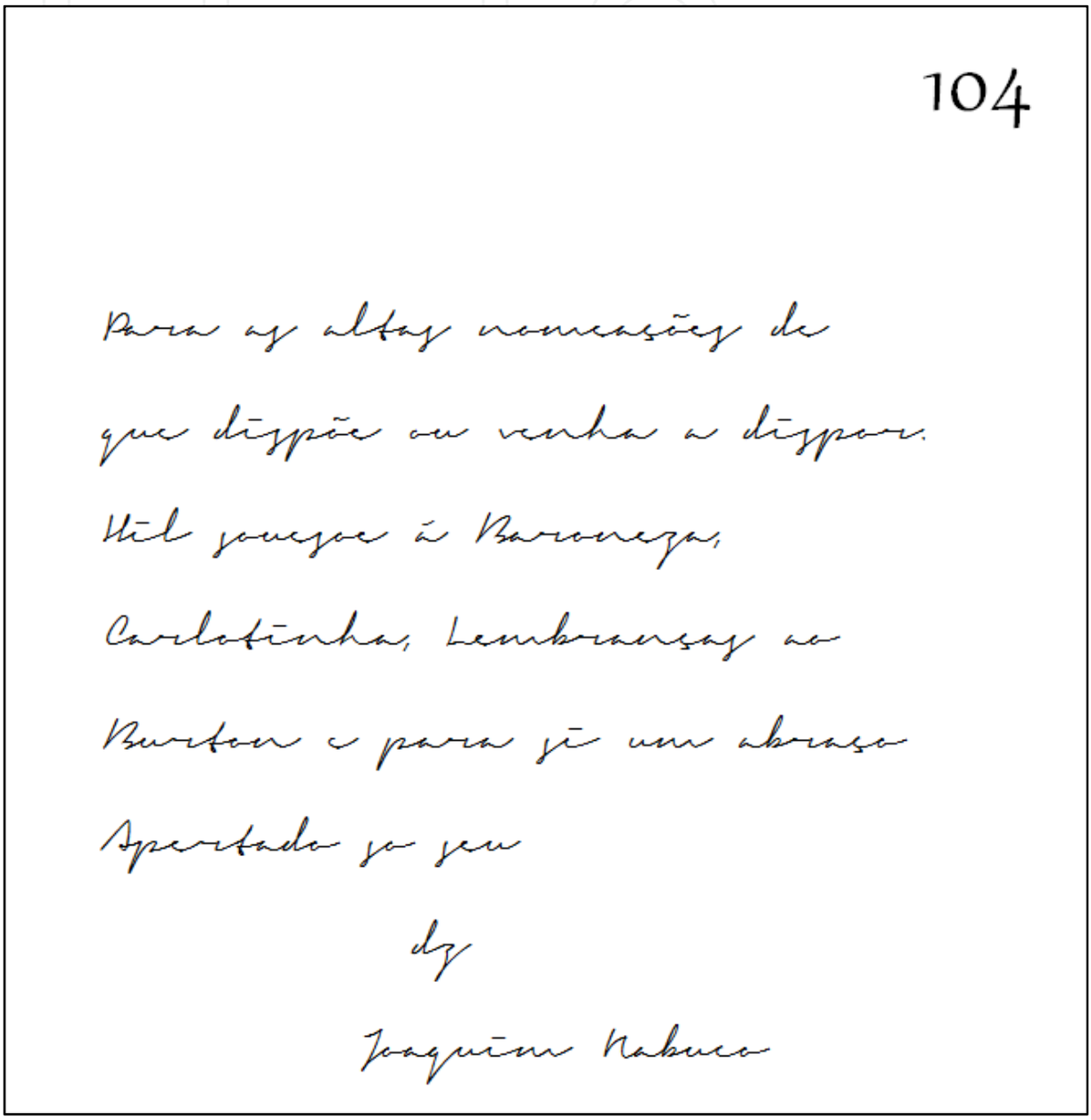


Figure 6. Image of Figure 5 after vectorization and deformation to make a font case matching to the original document after skeletonization and dilatation (Figure 4).

The comparison of the images of Figure 2 and Figure 5 shows several small differences, but there is a mapping path between each letter in the original text (ASCII character) and a "font" that resembles the author calligraphic pattern, which allows the automatic generation of a dictionary of patterns to be used as a training set for the recognizer.

The comparison of the images of Figure 2 and Figure 5 shows several small differences, but there is a mapping path between each letter in the original text (ASCII character) and a "font" that resembles the author calligraphic pattern, which allows the automatic generation of a dictionary of patterns to be used as a training set for the recognizer.

A MLP [8] and two SOM [10] fuzzy classifiers were used in parallel and the majority vote is taken for the transcription of the 25 letters in the document test set, totaling 2,115 words (with at least three letters), both trained with the same dictionary of synthesized words. The result obtained was of 61% (1,294 words) correctly transcribed and 17% (364 words) mismatched into (incorrect) valid words. Testing the whole set of fifty letters (3,584 words), that include the 25 letters used to develop the training set the results were of 67% words correctly transcribed and 15% of "false-positive" words . The result of the classifier applied to the remaining 25 letters yielded a precision and recall of 72%. Table 1 shows the significance of the recognition rate reached may be seen if one attempts to automatically recognize the document in Figure 2 with the classifier trained using the approach presented here and three of the best OCR softwares available today in the market: the Abby FineReader version 12 [19], Omnipage [23] and OCRopus 0.3.1 (alpha3) [22] that calls Tesseract.

Table 1 witnesses the suitability of the melhod proposed here. It is interesting to notice that even the human reader does not know what Joaquim Nabuco meant with the symbol (?) just before his signature. The transcription automatically made using the methodology proposed here may be considered very successful, overall if compared with the transcriptions obtained by the commercial OCRs tested (Tesseract produced no output at all). One interesting fact to observe is that although the grammatically correct accent in the third line of the text is "à", Nabuco´s writing was very calligraphically "imprecise" and looks as "á", as automatically transcribed. One may not consider that an error or even that he mispelled the lexeme "à", because the "á" in isolation does not exist in Portuguese. The addition of a dictionary may solve such a problem as well as some other as for instance the transcribed word "Hil" does not exist in Portuguese and the only possible valid candidate is the correct word "Mil" (one thousand).

3.2. Word recognition in death certificates

Death certificates provide important data such as *causa mortis*, age of death, birth and death places, parental information, etc. Such information may be used to analyze not only what caused the death of the person, but also a large number of demographic information such as internal migration, the relation of death cause with marital status, sex, profession, etc.

Images were acquired by The Family Search International Institute using a camera-based platform.

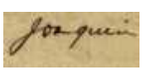
104 Para as altas nomeações de que dispõe ou venha a dispor. Mil saudades à Baroneza, Carlotinha, Lembranças ao Burton e para si um abraço Apertado do seu ??? Joaquim Nabuco		104 Para as altas nomeações de que dispõe ou venha a dispor. Hil souesoe á Baroneza, Carlotinha, Lembranças ao Burton e para si um abraço Apertado do seu dz Joaquim Nabuco
Human transcribed text		**Proposed Method**
1*^	-* * ‘ ^ ^^“f* ^CjL-íU As À-«-$tjt_Jt-C 	
Omnipage	**Abby FineReader Professional Edition 12**	**OCRopus/Tesseract**

Table 1. Human transcribed text of the document in Figure 2 and the automatic transcriptions by the Proposed Method, Omnipage, Abby FineReader and Tesseract.

Thanatos [1] is a platform designed to extract information from the Death Certificate Records in Pernambuco (Brazil), a collection of “books” kept by the local authorities from the 16th century onwards. The current phase of the Thanatos project focuses on the books from the 19th century. During such period, registration books were pre-printed with blank spaces to be filled in by the notary, as shown in Figure 7. Pre-processing is performed to remove noisy borders using the algorithm described in reference [6] incorporated in the

HistDoc Plarform as this step influences all the result of the other subsequent algorithms, the result of which is shown in Figure 8. Image processing continues on the border-removed image (Figure 8) to make image-size (resolution) uniform, binarize, correct skew (using the algorithm Ávila and Lins [3], 2005 also implemented in HistDoc [13]), remove salt-and-pepper and clutter noises, and finally splitting an image in two images each of them corresponding to one death certificate as shown in Figure 9.

Notaries in Brazil are a concession of the State. They are a permanent position many people exercise throughout their lives. Thus, most record books are written by a single person, allowing one to use the strategy proposed here to train the classifier to recognize the content of the different fields. Masks are then applied to extract the content of each of the fields filled in by notaries to extract the content.

They are:

- Nº (Register number) – placed at the top of the left margin of the register. It conveys numerical information only. Example: Nº 19.945.
- Data (Date) – the date is written in words and the information is filled in three fields for day, month, and year in this sequence. Example: Aos vinte e três dias do mês de janeiro de mil novecentos e sessenta e seis (At the twenty three days of the month of January of one thousand nine hundred and sixty six).
- Nome do cartório (Notary name) – this field holds the name of the place where the notary office was found. Example: neste cartório da Encruzilhada (at this notary office at Encruzilhada).
- Município do Cartório (City of the notary office) – Example: município de Recife (at the city of Recife).
- Estado do Cartório (State of the notary office) - Example: Estado de Pernambuco (State of Pernambuco).
- Nome do Declarante (Name of declarer) – Name of who attended the office to inform the death. Example: compareceu Guilherme dos Santos (attended Guilherme dos Santos).
- Nome do Médico (Name of the Medical Doctor) – Name of the M.D. who checked the death. Example: exibindo um atestado de óbito firmado pelo doutor José Ricardo (showing a death declaration signed by doctor José Ricardo).
- Causa mortis – Specifies the reason of the death in the declaration from the M.D. Example: dando como causa da morte edema pulmonar, o qual fica arquivado (that states as cause of death lung edema, which is filed).

The first strategy reported in reference [1] for information recognition in the Thanatos platform was to transcribe the fields using the commercial OCR tool ABBYY FineReader 12 Professional Editor [19]. The results obtained were zero correct recognition for all fields, including even the numerical ones. Such disappointing results forced the development of a recognition tool for the Thanatos platform based in the approach in reference [17] that makes use of a set of geometrical and perceptual features extracted from "zoning" the image.

"Zoning" may be seen as splitting a complex pattern in several simpler ones [18] [11] [7]. The original Thanatos strategy used dictionaries to analyze the possible "answers" to the blank fields. The original results of tests performed with 300 death certificates extracted from the same book of death records [1] were already considered reasonable and are shown in the first column of Table 2.

The adoption of the strategy presented here to generate the features of the writer through the modification of a cursive type font text was adopted. The list of all cities and places (villages, neighborhoods, etc) in the state of Pernambuco was collected from IBGE (the Brazilian Geographic and Statistical Institute) a social science research institute responsible for demographic and economic statistics and data collection in Brazil. Another list of family names was also generated having as basis the local phone directory. Those lists were "typeset" using the synthetic set of features extracted and then used to train the classifiers. The results obtained adopting this strategy is presented in the New column in Table 2. It is important to stress that the same parallel architecture (MLP + 2 SOM) fuzzy classifiers with majority vote was used in both cases, only with different training sets.

Field	Thanatos	New
Name of Notary	98.0%	98.5%
City of the Notary	71.0%	94.0%
State of the Notary	98.0%	100.0%
Place of death	31.0%	73.0%
Numbers in writing - Time of obit	69.0%	91.0%
Numbers in writing - Date of death	69.0%	91.0%
Numbers in writing - Date of birth	69.0%	91.0%
Color of skin	100.0%	100.0%
Marital status	100.0%	100.0%

Table 2. Recognition rate for non-numerical fields in 300 certificates.

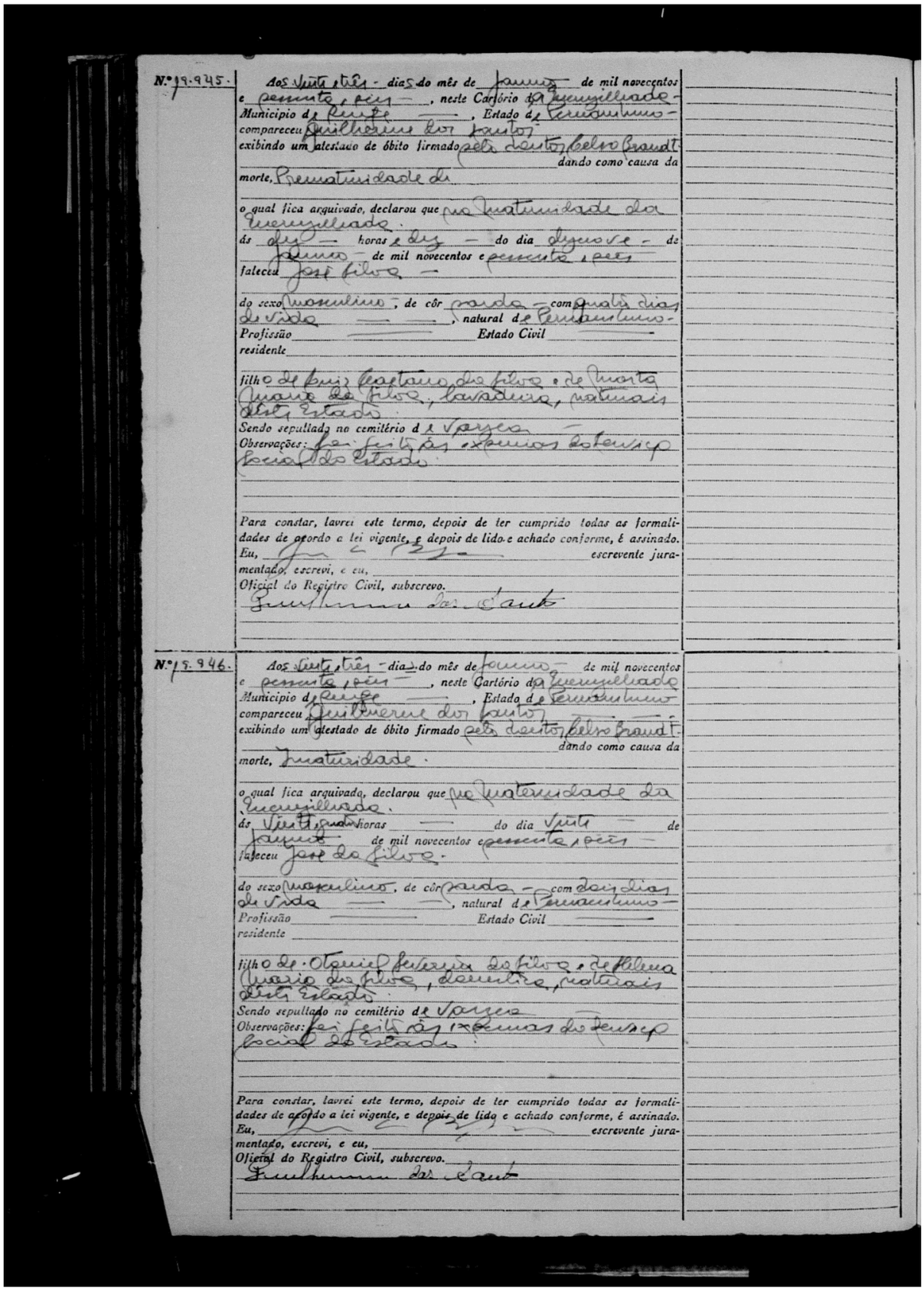


Figure 7. Original image from a book of printed forms of death certificates in Pernambuco (Brazil) – 1966.

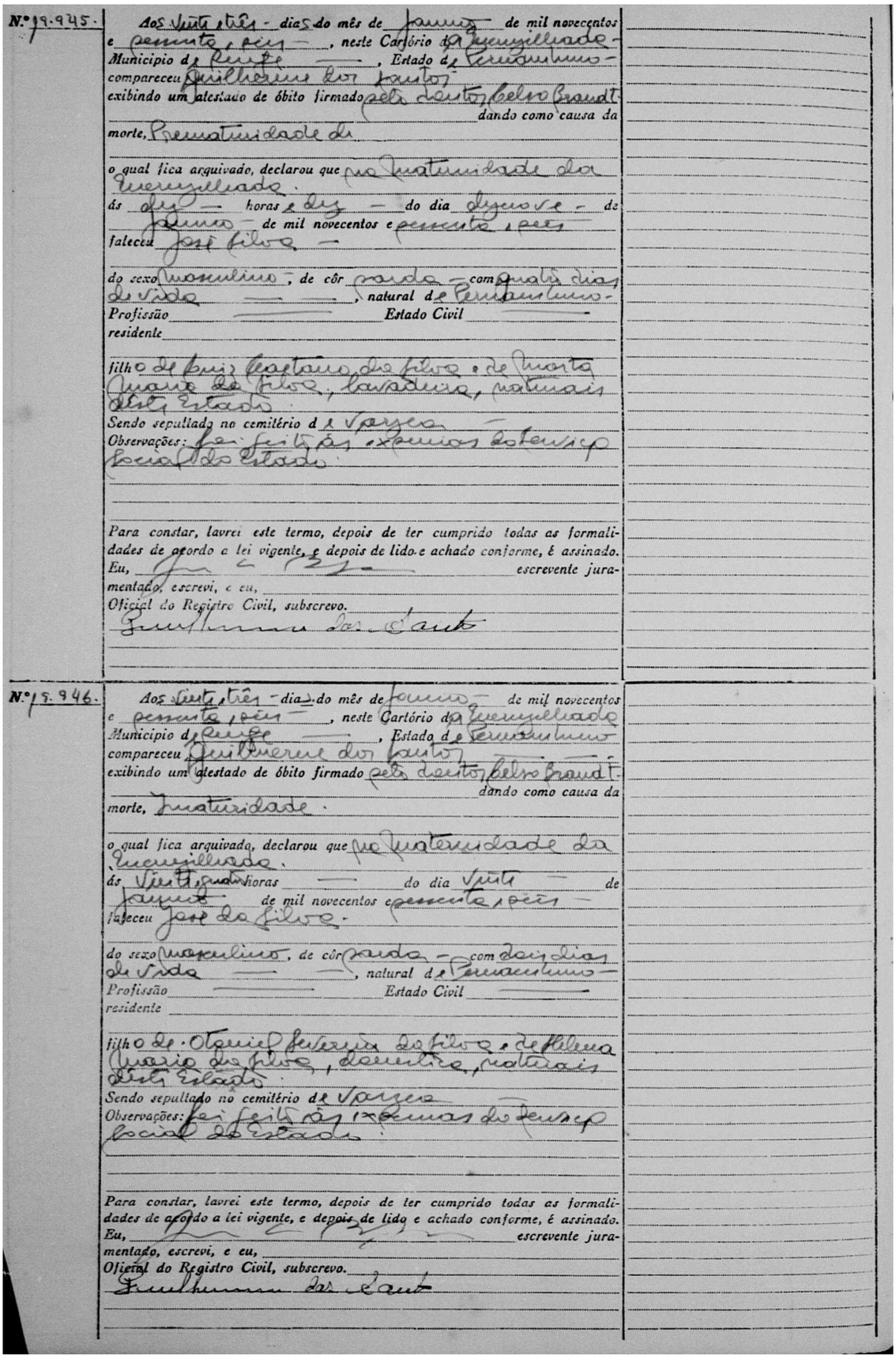


Figure 8. Filtered version of the image in Figure 7.

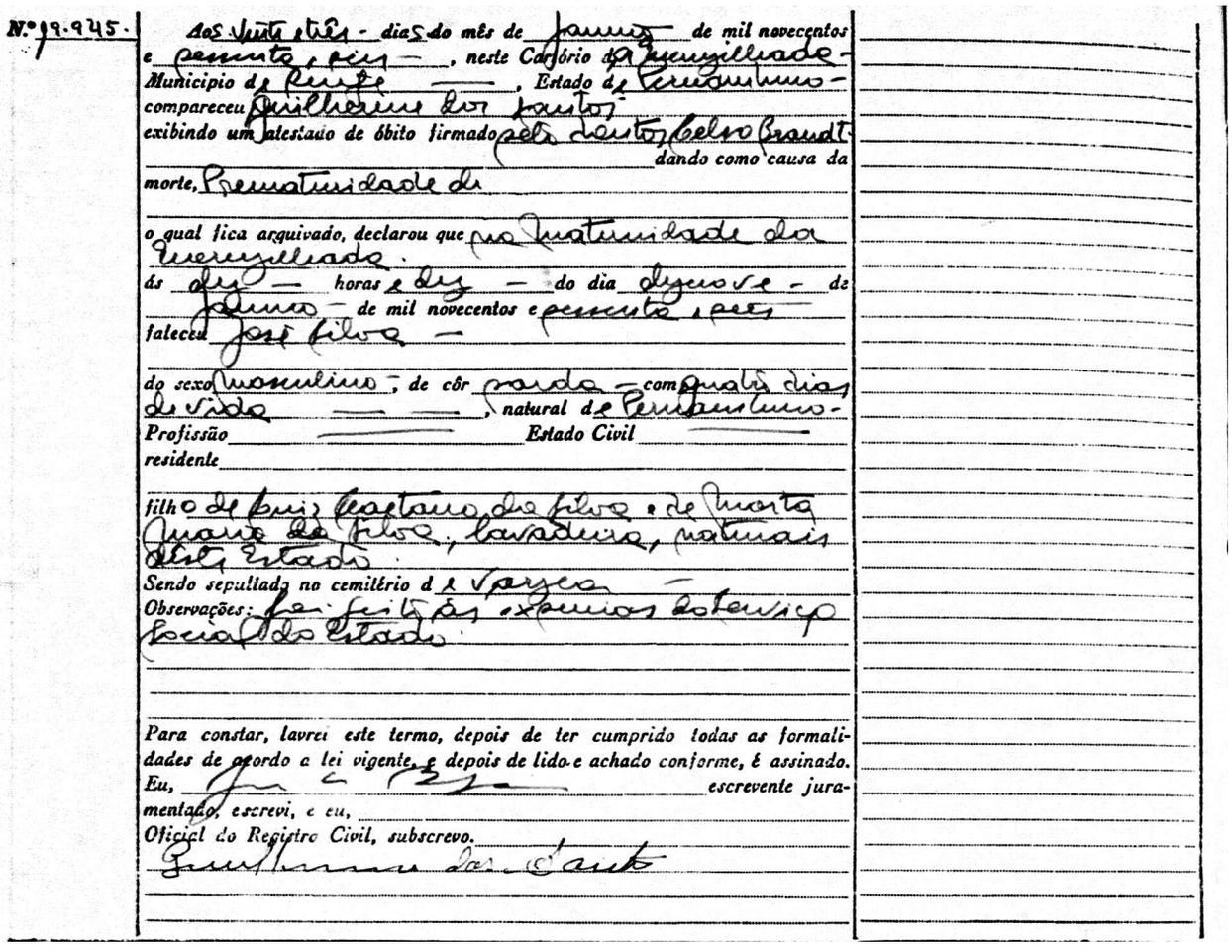


Figure 9. Monochromatic version of Death Certificate after filtering and splitting the image in Figure 8.

The column Thanatos refers to the results obtained in reference [1], while **New** presents the results of the strategy presented in this paper. Table 2 shows that the new strategy presented here presented either no loss or gains in the recognition rate of all fields recognized in relation to the results presented in reference [1]. In the case of the field "Place of death" the increase in recognition rate reached 42%.

4. Conclusions and lines for further work

Handwritten recognition to gain any degree of success either chooses a limited dictionary and allows a large number of writers or widens the vocabulary and largely restricts the numbers of writers. In both cases, the choice of the training set is of central importance for the success of any classification strategy and must be representative of the whole "universe" one wants to correctly recognize. This paper presents a new way of automatically generating the training set for the recognition of a large set of words written by a single user. It has as starting point different sets cursive type fonts, which are modified and compared to the original writing to "match their features". Once the "matching path" is found it is applied to a large dictionary in that encompasses the vocabulary of the document, generating the training set to be used for the whole batch of documents to be transcribed.

The strategy presented here was used with success in two sets of documents. In the case of the transcription of the handwritten letters in the bequest of Joaquim Nabuco it reached the correct rate of 67% transcribed words (of more than three letters), a result that may be considered successful at least for keyword indexing of such historical documents. In the case of death certificates of the Thanatos project, whose vocabulary is far more restricted the results presented either no loss or gains in the recognition rate of all fields recognized in relation to the previous results, reaching an average of 93.79% correct field recognition.

The statistical data collected inter character and inter word spacing, line and character skew, inter line separation were not used to enrich the generation of entries in the dictionary of the training set. Its use is left as a possibility for further work.

Author details

Gabriel Pereira e Silva and Rafael Dueire Lins
Universidade Federal de Pernambuco, Brazil

Acknowledgement

The authors are grateful to the organizers of the Fontspace site for setting such a useful site, fundamental for the development of this work. The authors also thank the Family Search International Institute for the initiative of digitizing the death certificate records of Pernambuco (Brazil) and to Tribunal de Justiça de Pernambuco (TJPE) to allow the use of such data for research purposes.

Research presented here is partly sponsored by CNPq-Conselho Nacional de Pesquisas e Desenvolvimento Tecnológico, Brazilian Government.

5. References

[1] A. Almeida, R.D.Lins, and G.F. Pereira e Silva. Thanatos. Automatically Retrieving Information from Death Certificates in Brazil. Proceedings of the 2011 Workshop on Historical Document Imaging and Processing, pp. 146-153, ACM Press, 2011.

[2] A. I. de S. L. Andrade, C. L. de S. L. Rêgo, T. C. de S. Dantas, Catálogo da Correspondência de Joaquim Nabuco 1903-1906, volume I 1865-1884, volume II 1885-1889, volume III 1890-1910, Editora Massangana, ISBN 857019126X, 1980. (Available at: www.fundaj.gov.br/geral/2010anojn/catalogo_nabuco_v2.pdf)

[3] B. T. Ávila and R. D. Lins. A Fast Orientation and Skew Detection Algorithm for Monochromatic Document Images. 2005 ACM International Conference on Document Engineering, p.118 - 126. ACM Press, 2005.

[4] L. Bethell, J. M. De Carvalho. Joaquim Nabuco, British Abolitionists, and the End of Slavery in Brazil: Correspondence 1880-1905, Institute for the Studies of the Americas, 2009. ISBN-13: 978-1900039956.

[5] M. Bulacu and L. Schomaker. Text-independent writer identification and verfication using textural and allographic features. IEEE Trans. on PAMI, 29(4):701.717, April 2007.

[6] A. de A. Formiga and R. D. Lins. Efficient Removal of Noisy Borders of Monochromatic Documents. International Conference on Image Analysis and Recognition, 2009, LNCS v.5627. p.158 – 167, Springer Verlag, 2009.

[7] C. O. A. Freitas, L.S. Oliveira, S.B.K. Aires, F. Bortolozzi, Zoning and metaclasses for character recognition. ACM–SAC 2007. P. 632-636, 2007.

[8] S. Haykin, Neural Networks: A Comprehensive Foundation. Hardcover (2nd Edition), 1998.

[9] M. Hu, "Visual pattern recognition by moment invariants", IEEE Transactions on Information Theory, 8(2):179-187, 1962.

[10] T. Kohonen, Self-Organizing Maps, Springer Series in Information Sciences,Springer, second edition, vol. 30, 1997.

[11] Z.C. Li, C.Y. Suen, J. Guo, A Regional Decomposition Method for Recognizing Handprinted Characters, IEEE Transactions on Systems, Man, and Cybernetics, N. 25, p. 998-1010, 1995.

[12] R.D.Lins. Nabuco - Two Decades of Document Processing in Latin America, Journal of Universal Computer Science, v. 17(1), pp. 151-161, 2011.

[13] R.D.Lins, G.F. Pereira e Silva, A.de A. Formiga. HistDoc v. 2.0: enhancing a platform to process historical documents. Proceedings of the 2011 Workshop on Historical Document Imaging and Processing, pp. 169-176, ACM Press, 2011.

[14] C. Liu, Y. Liu, and R. Dai, "Preprocessing and statistical/structural feature extraction for handwritten numeral recognition", Progress of Handwriting Recognition, A.C. Downton and S. Impedovo eds., World Scientific, 1997.

[15] C. Liu, K. Nakashima, H. Sako, and H. Fujisawa, "Handwritten digit recognition: benchmarking of state-of-the-art techniques", Pattern Recognition, 36(10):2271-2285, 2003.

[16] L. Oliveira, R. Sabourin, F. Bortolozzi, and C. Suen, "Automatic recognition of handwritten numerical strings: A recognition and verification strategy", IEEE Trans. on Pattern Analysis and Machine Intelligence, 24(11):1438-1454, 2002.

[17] G. F. Pereira e Silva and R. D. Lins. An Automatic Method for Enhancing Character Recognition in Degraded Historical Documents. ICDAR 2011, Beijing, September, IEEE Press, 2011.

[18] C.Y. Suen, J. Guo, Z.C Li, Analysis and Recognition of Alphanumeric Handprints by parts, IEEE Transactions on Systems,Man, and Cybernetics, N. 24, p. 614-631, 1994.

[19] ABBYY FineReader 12 Professional Editor, http://finereader.abbyy.com/ , last visited on 13.04.2012.

[20] IMAGEJ. http://rsbweb.nih.gov/ij/ , last visited on 13.04.2012.

[21] http://www.fontspace.com/category/cursive?p=19 , last visited on 13.04.2012.

[22] OCRopus 0.3.1 (alpha3): http://code.google.com/p/ocropus/
[23] Nuance OminiPage Professional 16: http://www.nuance.com/for-individuals/by-product/omnipage/index.htm

IntechOpen